The Scamdemic Conspiracy Theory and Twitter's Failure to Moderate COVID-19 Misinformation

Rossi, Sippo

Document Version
Final published version

Published in:
Proceedings of the 55th Hawaii International Conference on System Sciences

DOI:
10125/79345

Publication date:
2022

License
CC BY-NC-ND

Citation for published version (APA):
Rossi, S. (2022). The Scamdemic Conspiracy Theory and Twitter's Failure to Moderate COVID-19 Misinformation. In T. Bui (Ed.), *Proceedings of the 55th Hawaii International Conference on System Sciences* (pp. 124-133). Hawaii International Conference on System Sciences (HICSS). https://doi.org/10125/79345

Link to publication in CBS Research Portal

The Scamdemic Conspiracy Theory and Twitter's Failure to Moderate COVID-19 Misinformation

Sippo Rossi
Copenhagen Business School
sr.digi@cbs.dk

Abstract

During the past few years, social media platforms have been criticized for reacting slowly to users distributing misinformation and potentially dangerous conspiracy theories. Despite policies that have been introduced to specifically curb such content, this paper demonstrates how conspiracy theorists have thrived on Twitter during the COVID-19 pandemic and managed to push vaccine and health related misinformation without getting banned. We examine a dataset of approximately 8200 tweets and 8500 Twitter users participating in discussions around the conspiracy term Scamdemic. Furthermore, a subset of active and influential accounts was identified and inspected more closely and followed for a two-month period. The findings suggest that while bots are a lesser evil than expected, a failure to moderate the non-bot accounts that spread harmful content is the primary problem, as only 12.7% of these malicious accounts were suspended even after having frequently violated Twitter's policies using easily identifiable conspiracy terminology.

1. Introduction

We may be living in a golden age of conspiracy theories [1, 2]. In everyday life, you can hear terms such as QAnon and Pizzagate, which previously belonged to the vocabulary of fringe groups but have increasingly been adopted by wider audiences and normalized as part of the vernacular by the news media and elected officials [3]. This "normalization" of conspiracy theories started well before the COVID-19 pandemic as several politicians, particularly in the United States, began promoting them during the 2016 and 2020 presidential elections. However, the coronavirus resulted in an eruption of new theories, ranging from 5G causing the virus [4] to the Great Reset that suggests the pandemic is being used as an excuse to take control of the world economy [5].

The role of the social media platforms in spreading conspiracy theories has been tied to their loose regulation as well as to conspiracy theorists taking advantage of their algorithms to amplify the spread of content, bringing it from the obscure corners of the internet to the mainstream feeds of the general public. Some evidence supports the idea that bad actors have also become better at exploiting vulnerabilities of the algorithms that power modern content recommendation systems [6].

Furthermore, recent studies have suggested that the conspiracy theories and misinformation related to the COVID-19 pandemic have been amplified by suspected bot accounts [7, 8]. For example, bots have been used to distribute conspiracy theories related to the pandemic alongside references to QAnon and the Great Awakening as well as to share links to other low credibility content and fake news sites [7]. Although the existence of bots in this context has been proven, the estimates of their prevalence range wildly.

To reduce the reach of conspiracy theories, social media sites such as Twitter, Facebook and YouTube have begun moderating content more aggressively, suspending accounts that spread misinformation and shutting down groups that are devoted to or helping spread conspiracy theories [9]. For instance, many social networking sites have attempted to remove links and references to the Plandemic, which was a viral video that promoted several conspiracy theories related to COVID-19 and has become a term used to refer to the pandemic as an orchestrated epidemic or hoax [10]. While some social media sites like Facebook have seen a decline in the number of interactions with content containing misinformation during the recent years, in Twitter interactions with such content have been steadily growing [11].

Moderating content on conspiracy theories is challenging. The propagators' methods have evolved quickly, adapting to restrictions, developing new terms and adding nuance to content that is harder to identify as misinformation. When searching on Twitter with the term Plandemic, the social media site redirects the search to the word "pandemic" and provides a link to official information about the COVID-19 pandemic.

URI: https://hdl.handle.net/10125/79345
978-0-9981331-5-7

But, when using the search term Scamdemic, Twitter does not attempt this redirection and users can even see tweets where the word or hashtag Plandemic is used in unison with the word Scamdemic. The difference between the words is minor, but the new term is unmistakably related and has been widely used without consequences. This is surprising considering Twitter's policy update which banned sharing conspiracy theories about the pandemic or COVID-19 vaccines [12].

Due to the possible negative effects on society and public health that conspiracy theories around the COVID-19 pandemic may have, it is important to evaluate the effectiveness of these content bans and to understand what is driving the conspiracy theories so that they can be addressed accordingly. The possible involvement of bots in the COVID-19 conspiracy theory discussions further complicates the study of the topic as well as the analysis of policies, as moderation of computer-generated content is far less ambiguous than the moderation of content made by humans. This is due to the use of bots for manipulation being clearly banned and the removal of such accounts will not result in dissatisfied users, whereas moderating genuine human accounts can lead to accusations of stifling free speech as well as users fleeing to competing social networking sites. Analysis of the level of bot involvement is thus needed as it will increase our understanding of whether the primary issue is genuine or inauthentic accounts. Furthermore, the policy recommendation will vary depending on what types of accounts are the primary source of misinformation.

At the time of writing only few publications addressed the effectiveness of Twitter's misleading information policy that was adjusted multiple times during the years 2020 and 2021 to reduce COVID-19 related misinformation. The policy changes adjusted the criteria for suspensions and content removal and introduced a strike system for accounts, where repeatedly tweeting content containing misinformation would eventually lead to permanent suspension [12]. To address this research gap, and to provide a basis for future research, an exploratory study was conducted on a sample of COVID-19 conspiracy theory tweets that are using the conspiracy term "Scamdemic", which is commonly used both as a hashtag or as a keyword in the tweet to signal that the pandemic is a conspiracy or hoax.

The goal of this paper is two-fold. Firstly, it will investigate who are using the Scamdemic term on Twitter in order to determine whether the conspiracy theories are being pushed by coordinated attacks by bots and trolls or organically by users that believe in the conspiracy theories. Based on the findings on what type of accounts and content are evading the bans, the effectiveness and level of enforcement of current policies could be evaluated and policy changes suggested. Secondly, it will evaluate how well Twitter's COVID-19 misleading information policy is enforced using the tweets containing the word "Scamdemic" as a case example.

2. Related research

The literature review is divided into three parts. The first part will summarize research related to the spread of misinformation in social media and explain what is currently known of the characteristics of COVID-19 conspiracy content on Twitter. The second part will discuss what is known of bots and their ability to influence discussions on Twitter and provides a motivation for the method that will be used for bot detection in this study. Lastly, articles that are methodologically similar to this paper and use network analysis to study misinformation on Twitter will be reviewed. Overall, the goal is to establish and describe what has previously been observed in misinformation research and to justify the design choices outlined in the methodology section.

2.1 Misinformation and conspiracy theories on social media

The spread of conspiracy theories and misinformation has been widely studied [13] and the role of social media in distributing such content is a well-known issue [7]. New conspiracy theories and adaptations of earlier ones have appeared during the COVID-19 pandemic [14] and quickly reached wide audiences through social media platforms [4]. The conspiracy theory on the "Plandemic" which trended in multiple social media sites as a result of a viral video, is an example of COVID-19 influencing and modifying existing conspiracy theories related to vaccines [10].

One distinguishable characteristic of tweets involving conspiracy theories is that certain groups of hashtags and keywords are prevalent in them. For example, 5G is commonly mentioned due to the popularity of the related conspiracy theory that suggests the technology's emergence is linked to the disease [4]. Plandemic tweets also often include hashtags or mentions of other indirectly related or unrelated conspiracy theories and common examples include QAnon and the Deep State [7, 10]. Furthermore, many tweets containing misinformation include links to both YouTube videos as well as fake news websites [4].

Social engagement metrics such as likes and retweets have been shown to increase the susceptibility of users to posts containing misinformation [15]. This suggests that the virality of conspiratorial content and other misinformation is a threat precisely because people are unlikely to critically evaluate the source. Based on this, posts made by influential accounts that are retweeted and or liked in large numbers are a bigger threat than those made by less popular accounts, thus suggesting the focus of the analysis should be on them.

Influential accounts are not the only issue behind the propagation of misinformation, as the design of the platforms and the way they promote content is a major part of the problem as well. One of the theories that is used to both describe behavior in online social networks as well as to support or oppose restrictive policies is the concept of “echo chambers” or alternatively “filter bubbles” [16, 17]. These echo chambers are formed as a result of recommendation systems that aim to maximize interaction by providing users of the social networking site with content such as tweets that matches their views and suggestions on which accounts to follow that share similar content [16]. The danger according to this theory is that an individual will be eventually exposed mainly to material that aligns with their world view giving a false sense of unanimity while only a small minority of individuals supports the belief. Due to not seeing material that challenges for example the conspiracy theories, the individuals become more entrenched in their bubble or echo chamber [17]. However, too aggressive moderation of content or the banning of entire communities is a risk as it may result in the users abandoning the platform and moving to an alternative social networking site that may further increase the divide and drive individuals to the fringe.

2.2 The role of Twitter bots in the distribution of misinformation

When using the term bot or bot account, this paper refers to basic spambots as well as social bots that are either fully or partially automated and engaging in distributing controversial content without self-identifying as non-human. This is based on the definition given by Ferrara et al. [18].

Many papers have discussed the role of bots in distributing fake news and misinformation [6, 19]. It is argued that at least their role in distributing content from low-credibility sources is disproportionately big [19]. The percentage of bots in the entire Twitter population has been estimated to be around 10% - 20%. However, in the case of accounts pushing the United States to reopen the country and to reduce COVID-19 restrictions it has been up to 50% [8]. Furthermore, there is evidence of social bots that are interacting predominantly with COVID-19 content, suggesting that their purpose is to spread or amplify misinformation related to the pandemic [7].

One of the negative effects of bots, which further demonstrates their potential in the context of disseminating misinformation, is their assumed role in strengthening the spiral of silence [20]. The spiral of silence theory suggests that individuals monitor and attempt to understand the general opinion on a given topic and if they perceive themselves to be supporting the stance of the minority, they are likely to refrain from expressing their opinion [21, 16]. This ultimately affects other people’s perception of the topic and can lead to a setting where a silent majority accepts that the opposing view is the prevailing opinion of the population, while in fact it is supported by a vocal minority [21]. One recent study suggests that even a relatively small percentage of bots can affect online discussion and tip the perceived public opinion [20].

A major issue in studies that investigate the role of bot accounts in the spread of misinformation is the difficulty of reliably detecting modern social bots. Recent papers focusing on Twitter bot detection rely increasingly on machine learning [22, 23, 24] and ensemble methods combining multiple classifiers due to the level of sophistication of bots as well as hybridization where both humans and programs control the accounts [25]. One particularly widely used example has been the Botometer (or originally BotOrNot), which has been featured in many of the most cited publications on social bots [26, 27]. However, fully automated bot detection may not be realistic [25, 26] because the results of such techniques have been shown to vary with new datasets. Therefore, relying on existing tools such as the Botometer alone and drawing conclusions without critical qualitative inspection appears no longer sufficient and thus in this paper a hybrid approach combining algorithmic and qualitative labeling is employed.

2.3 Networks on Twitter

Social network analysis has been widely used to study social media [28], how information spreads in networks [29] and which accounts are most influential in facilitating information spread [30]. Network models based on Twitter data can be built in multiple ways with the simplest examples being models where connections represent accounts following each other or mentioning each other in tweets. The networks can also represent relationships between content that is being shared such as tweets containing the same hashtag, and in that case the nodes can be the hashtags.

Research has shown that in the case of misinformation, unverified accounts that do not belong to any well-known public figures influence the spread of conspiracy theories [31, 32]. However, the way in which these influential accounts are defined varies a lot and influence can be measured in many ways. For example, when defining influence algorithmically, the most influential accounts can be those that are surrounded by highly retweeted accounts who commonly share the content of the less well-known account [31]. Simpler approaches rely on using different metrics related to the Twitter accounts such as the number of followers [30] and betweenness centrality [4].

3. Methodology

3.1 Data collection

The dataset consists of Twitter usernames, tweets and a mapping of the relationships between the different objects, which will be described in more detail under the network analysis section. The data was collected using Twarc, a Python library for accessing and retrieving data from the Twitter API.

The data contains 8263 tweets and 8540 users interacting with or being related to these tweets. The data was gathered from the Twitter API with the tweets/search command and search term "scamdemic". Users are considered related to a tweet if the tweet mentions the user, retweets or quotes the user or if it is a reply to a tweet made by the user. The dataset contains tweets posted during a one-week period starting on the 8th of March and ending on the 15th of March 2021. The time was chosen based on Twitter having updated their COVID-19 misinformation policy at the beginning of the month. The script used to collect and process the data is based on a tool described in [33].

3.2 Network analysis

The data was mapped so that two separate network graphs can be created. The first one is labeled the account-interaction network, which is a weighted directed network where nodes are accounts while the edges represent interactions towards other accounts with tweets. Weights are determined by how many times during the analysis period an account interacted with the other account by for example retweeting or mentioning them. The second is labeled the account-hashtag network and is a directed multimodal network where nodes are both hashtags as well as accounts and the edges indicate which hashtags an account interacted with. From now on, the first network will be referred to as the account-interaction network and the latter as the account-hashtag network. The networked data was analyzed both quantitatively, with standard network analysis metrics, as well as qualitatively by manually inspecting the most important nodes' Twitter profiles. Overall, the purpose of this network analysis was to determine how the average account using the Scamdemic word behaved, which hashtags were used together, and by whom, and to identify which accounts were most prominent in the network.

To make inferences on the effectiveness of Twitter's policies, a population of influential accounts were selected based on the three node characteristics: betweenness centrality, indegree and the outdegree. A high indegree indicates that the account is often referred to in other tweets, while a high outdegree would indicate a spammer whose content are likely to be seen by individuals searching with the right keywords. Lastly, users with a high betweenness centrality are the accounts that act as a bridge between communities and discussions. Figure 1 illustrates these different node characteristics with the teal "A" node representing a node which has a high indegree, while the red node "B" has a high outdegree and the yellow node "C" a high betweenness centrality as it acts as a link between the two communities around the node "A" and node "B".

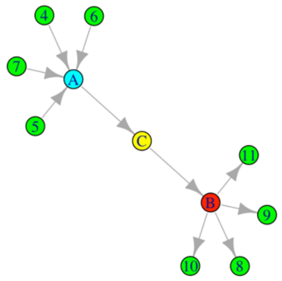

Figure 1. Node characteristics example

Heuristically, the 25 accounts with the highest betweenness centrality, in and outdegrees were determined to be influential, and consequently their activities reviewed twice during the two months following the collection of the dataset. As some influential nodes were in the top 25 of several characteristics, the final list of influential nodes consists of only 61 accounts. The rationale behind focusing on these accounts is that they should be among the first to be deleted due to their prominence assuming that they are in fact supporting the conspiracies and not attempting to debunk them.

3.3 Bot detection and classifying accounts

Several methods were used in conjunction to determine what types of accounts were participating in the discourse and if bots are amplifying the Scamdemic conspiracy. Firstly, the 61 most influential nodes were checked with the Botometer which provides a rating on the likelihood of the account being a bot rather than a classification. Secondly, manual inspection and coding was done to further validate the scores provided by the Botometer. This two-step classification of accounts should reduce the risk of misclassification tied to the Botometer's scores [26]. All influential accounts were checked even if the Botometer suggested that they are not suspicious.

In addition to labeling accounts as humans or bots, the manual inspection was used to bin the accounts into overlapping categories. The labels for these categories are conspiracy theorists, spammer, antivax, celebrity and non-believer. Conspiracy theorists are accounts that seemed to authentically believe and participate in the discussions. Spammers are accounts that solely push content through liking or retweeting. Antivaxxers are a subset of conspiracy theorists, mainly engaging with content questioning the safety of COVID-19 vaccines. Lastly, celebrity indicated prominent politicians and non-believers accounts that are participating in the discussions in order to debunk conspiracies. This manual inspection was conducted twice. First, a month after the collection of the dataset and a second time a month later to see on both occasions which accounts had been banned during the monitoring period and to follow-up on whether the coding was still accurate.

Figure 2 shows the account-interaction network with influential human accounts being marked as black, and influential nodes suspected of being bots as red. Due to the metrics used to determine influence, the influential nodes are mostly in the center of a cluster of accounts or acting as a link between several clusters.

3.4 Limitations

Accessing Twitter's API with Twarc does not guarantee that all tweets related to the search term are collected. This is due to the tool not supporting Twitter's academic product track's full-archive search. However, even small samples instead of full datasets have been successfully used in previous studies [4] and especially considering the relative niche status of the Scamdemic, a low volume of tweets can be expected. Furthermore, the dataset is small when compared to typical Twitter studies, but for the purposes of demonstrating how individual influential accounts can avoid bans while repeatedly posting content that is against the rules, it should be sufficient.

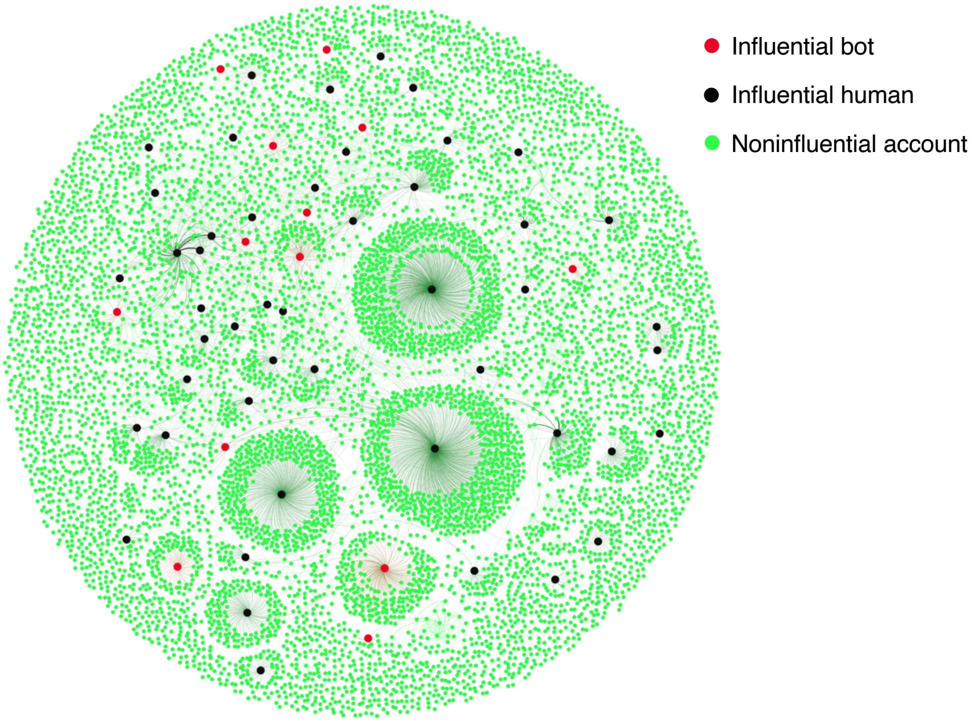

Figure 2. The account-interaction network with influential nodes highlighted

4. Findings

4.1 Accounts

The account-interaction network which consisted of all the 8540 accounts in the discussions is very sparse. Most nodes are peripheral or separate from the main network with 83% of the nodes having an outdegree between 0 and 1, while 90% of the nodes have an indegree between 0 and 1, meaning that they have interacted with another account or been mentioned in a tweet 0-1 times. In all three previously described network characteristics that were used to define the influential accounts; the indegree, outdegree and betweenness centrality, the top 10 to 25 accounts are distinguishable from the rest by having values that are several hundred or even thousand percent higher than the mean.

Figure 3 represents the account-interaction network that was created using Gephi with the Force Atlas 2 layout algorithm. The different colors represent communities that were identified based on the modularity. There are several influential nodes that have large communities of accounts interacting with them while most are in hardly visible small clusters of 2-3 accounts.

From the list of 61 influential accounts, only five were identified as public organizations, well-known individuals or politicians that commented on the conspiracies or who were mentioned in the discussions. Furthermore, only one of these five accounts was a supporter of COVID-19 conspiracy theories. Additionally, in the case of one of the participants, it was unclear whether they were debunking or promoting the conspiracies.

This is in line with previous research which suggests that with misinformation, most of the influential accounts are not verified users or public figures. The remaining 55 accounts were a mix of trolls, bots and users assumed to be authentic conspiracy theorists or believers and thus can be referred to as malicious accounts. Only six accounts had been banned after a month and three had been renamed and thus were no longer characterizable. On the second inspection two months after the data was collected, one additional account had been suspended and two more renamed making them untraceable.

Defining the exact type of the malicious accounts proved to be difficult as their goals were not clear in most cases. By looking at the profile descriptions and content that they tweeted and retweeted, in some cases the motives as well as their assumed country of origin

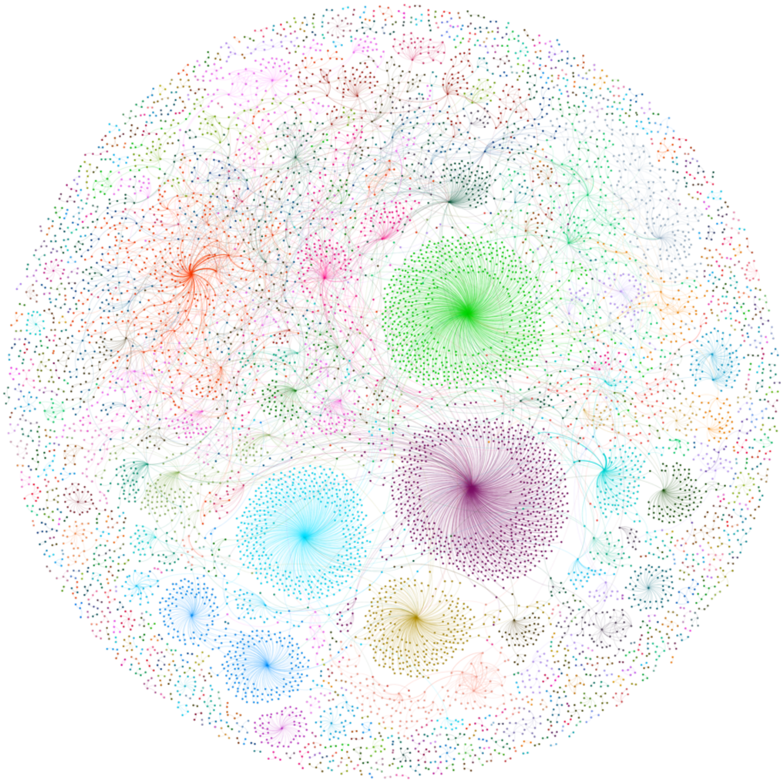

Figure 3. The account-interaction network clustered into communities

were identifiable. Surprisingly, over 40 percent of the influential accounts were interacting with COVID-19 content originating or related to Great Britain and British politics, which suggests that the word Scamdemic is popular in the British Twitter conspiracy theory circles, or that the sample was taken during a time in which the word was trending in the United Kingdom.

According to the Botometer, from the list of most influential accounts only five were given a score above 4 on a scale of 1-5 by the Botometer. The universal, or language independent score was used as not all accounts were tweeting in English. Two of these assumed bot accounts were banned at the time of writing and one was manually reclassified as a human on closer inspection. During the manual labeling, thirteen accounts were labeled as likely to be bots based on their behavior, which usually included mass retweeting and spamming of hashtags and mentions. The thirteen suspected bot accounts included all except one of the five accounts labeled by the Botometer as highly likely of being a bot. This would indicate that the qualitative labeling was more aggressive than the Botometer, which tends to be conservative with its estimates.

Overall, these thirteen suspected bot accounts represent a quarter of the influential accounts which is on the high end of the assumed share of bot accounts on Twitter, but on the low end of the estimates of the analyses that looked into the share of bots in COVID-19 misinformation tweets [6, 8].

4.2 The content

The content analysis focused on the different hashtags that were used since they play an important role in making a particular topic recognizable and easy to find on social networking sites such as Twitter. The capitalization was removed in order to combine some of the otherwise identical hashtags, such as Scamdemic and scamdemic which were initially treated as unique hashtags. A total of 3127 hashtags were used in the 8263 tweets and the ten most used ones represent 34.9% of all hashtags.

Figure 4 shows the account-hashtag network, where nodes are a mix of hashtags as well as the accounts that used them in their tweets. The communities are less distinct as in the previously discussed account-interaction network, which is due to the common practice of including multiple hashtags in posts. Several somewhat separate communities can be seen at the edges of the network, such as the pink cluster of non-English hashtags on the left side of the graph and calls to participate in rallies in the gray cluster at the top.

Unsurprisingly, the most frequently used hashtag was #scamdemic which was used over 500 times in the dataset, followed by the over 200 mentions of #covid19

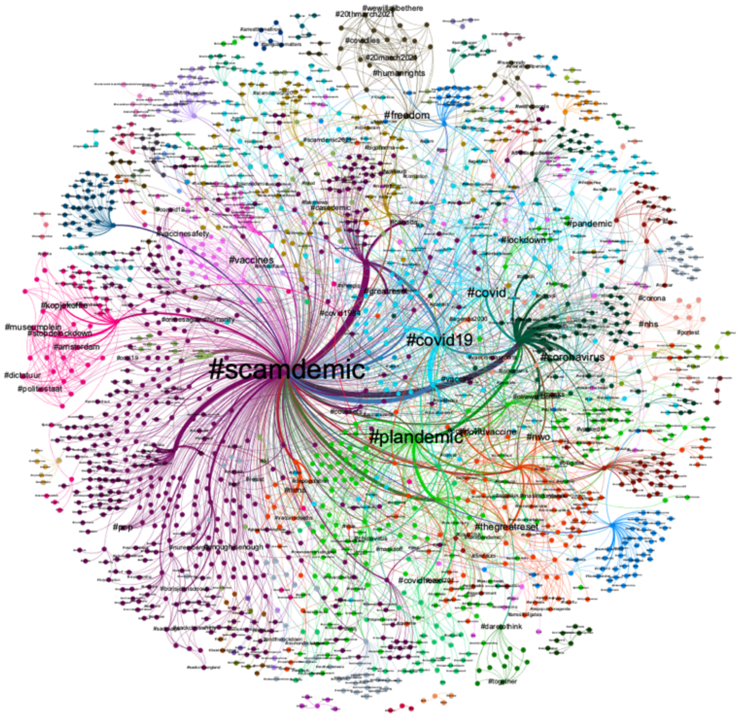

Figure 4. The account-hashtag network

in various ways of writing, which were merged with fuzzy matching. At third place was #plandemic which had been used over 100 times despite being a particularly scrutinized word. Other much used hashtags included the popular Great Reset (#thegreatreset) and New World Order (#nwo) conspiracy theories, as well as a large variety of different references to the COVID-19 vaccine. The table below shows the top ten hashtags, which includes generic pandemic related words such as lockdown and vaccines in addition to the terms linked to conspiracy theories. Most hashtags are in English, although German and other minor European languages were used in small numbers as well. Table 1 shows the ten most popular hashtags and how many times they were used in the dataset.

Table 1. Most popular hashtags

Hashtag	Count
#scamdemic	524
#covid19	220
#plandemic	114
#thegreatreset	41
#nwo	35
#coronavirus	31
#freedom	28
#vaccines	24
#lockdown	21
#covidvaccine	19

5. Discussion

5.1 Implications

One of the main objectives of the study was to determine the nature of the accounts that were participating in the distribution of the Scamdemic conspiracy term by looking closely at a sample of 61 highly active and influential accounts. Furthermore, by following these influential accounts for a duration of two months the research aims to highlight the lack of moderation and enforcement of Twitter's policies against misinformation. The design of the study makes it difficult to draw conclusion on the implications that the findings have on existing theories used in misinformation research. During the qualitative inspection of the influential accounts, the lack of critical comments against the COVID-19 conspiracy theories can however suggest that the active participants are within an echo chamber and or that the spiral of silence is making it difficult for the participants to voice critical comments, but this will be verified with more thorough analysis during future studies. Thus, the discussion in this paper will be centered on the empirical evidence and based on the key implications, a critical commentary on the current status is provided. Lastly, recommendations on how to adjust Twitter's misleading information policy are given.

Interestingly, a majority of the influential accounts that are using the Scamdemic word and participating in the spread of other related conspiracy theories, seem to be legitimate users rather than bots. Moreover, most of the suspected bot accounts were merely retweeting conspiracy theories constantly without producing any original tweets, indicating that they are operating with crude scripts rather than more sophisticated programs found in modern social bots. Of the 55 accounts defined as malicious and influential the rate at which they were banned is surprisingly low at 12.7%, with only 6 bans during the first month and one additional ban after two months. Previous research has focused predominantly on how modern misinformation is spread by advanced social bots and coordination but based on the sample used in this study, both the bots and humans could merely continuously retweet and post malicious content without consequences. Therefore, the level of sophistication of the accounts avoiding suspension is likely lower than previously assumed.

In order to analyze Twitter's moderation and how well they follow their new policy, we looked at the content produced and shared by the influential accounts. Content wise it seems that using indirect or novel words and hashtags to avoid suspension is not needed on Twitter. This is based on the observation that using words and hashtags known to be associated with misinformation, or directly implying for example that the pandemic is a hoax (e.g., #plandemic and #scamdemic) are not being removed.

Only one instance of a tweet being flagged as against Twitter's rules was detected during the review of the influential accounts. Based on this, the content is not getting actively flagged and censored even in obvious cases. Considering that flagging a tweet rather than deleting it is a much lighter approach and is already employed by Twitter, it is questionable why it is not used more actively.

From the findings on the accounts and content that they engage with, Twitter's ability or interest to enforce its COVID-19 misinformation policy seems very weak. Almost 90% of the inspected accounts were openly tweeting or retweeting using easily identifiable words and hashtags related to popular conspiracy theories without getting suspended. Approximately a quarter of the influential accounts were also spreading anti-vaccine content, which is another topic when discussed in the context of COVID-19 that violates Twitter's misinformation policy. It is especially surprising how these accounts that are posting multiple types of content that should automatically raise alarms do not get removed or filtered from public searches. Interestingly even names and bios containing the word covid and mentions of Scamdemic or other conspiracy words had managed to not be suspended.

5.2 Policy recommendations

Lastly, two suggestions on how to mitigate the further spread of misinformation are provided. Due to the simplicity of the accounts involved, relatively basic changes to policy would reduce the visibility of the misinformation and conspiracy theories.

Firstly, more aggressively suspending accounts according to the current misinformation policy based on repeated use of known conspiracy theory terms is suggested. Particularly accounts involved in the distribution of conspiracies such the Plandemic as well as other vaccine related misinformation have a clear lexicon and should be targeted similarly as the accounts spreading the Plandemic are on Facebook, where suspensions are given more frequently. Considering that Twitter already attempts to filter content by requiring an additional click to access the tweets when querying with the search term Plandemic, it is clear that they are already capable of identifying the misinformation but abstaining from removing it. In other words, this recommendation simply suggests that Twitter should enforce its own current policies.

Secondly, considering that the most incriminating hashtags and vocabulary such as the Plandemic and Scamdemic are used by the malicious accounts to make content easy to find, filtering the tweets containing them from the search results would reduce their visibility even without the need of removing the content or associated accounts. This would also avoid false positives leading to bans of accounts that are not promoting conspiracies but in fact attempting to debunk them.

6. Conclusion

The different misinformation, fake news as well as conspiracies surrounding the COVID-19 pandemic have been studied from many angles despite of the recentness of the topic. The goal of this paper was to contribute to the understanding of what types of accounts are distributing the conspiracy theories and misinformation related to the pandemic, as well as demonstrate that Twitter is not highly successful at mitigating the spread of misinformation. The study found limited evidence of bot accounts dedicated to spreading misinformation related to the COVID-19 pandemic as the share of assumed bots when compared to human operated accounts was lower than expected. However, the findings were in line with previous research that cites humans as the most likely cause of misinformation spreading. Lastly, the study suggests that stricter enforcement is needed, and that the situation could be improved by merely removing or filtering content that contains certain keywords or hashtags such as #scamdemic and #plandemic.

This paper highlighted how it is possible for influential accounts to repeatedly share content that is against Twitter's policies during a short time without having the content removed or the associated accounts suspended. Future studies would benefit of having a longer monitoring period than the two months used for this study, as this could provide insights on whether in the long-term enforcement of the policies is more successful. Furthermore, by expanding the list of keywords and conducting the longitudinal study on a larger group of accounts, more inferences could be made on which type of behavior and terminology in tweets manages to evade suspension.

7. References

[1] D. Freeman and J. Freeman, "Are we entering a golden age of the conspiracy theory?," *The Guardian*, Mar. 28, 2017. [Online]. Available: https://www.theguardian.com/science/blog/2017/mar/28/are-we-entering-a-golden-age-of-the-conspiracy-theory

[2] Z. Stanton, "You're Living in the Golden Age of Conspiracy Theories," *Politico*, 2020.

[3] A. Willingham, "How the pandemic and politics gave us a golden age of conspiracy theories," *CNN*, Oct. 03, 2020. [Online]. Available: https://edition.cnn.com/2020/10/03/us/conspiracy-theories-why-origins-pandemic-politics-trnd/index.html.

[4] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. L. Seguí, "COVID-19 and the 5G conspiracy theory: Social network analysis of twitter data," *Journal of Medical Internet Research*, vol. 22, no. 5, pp. 1–9, 2020.

[5] J. Goodman and F. Carmichael, "The coronavirus pandemic 'Great Reset' theory and a false vaccine claim debunked," *BBC*, Nov. 22, 2020.

[6] D. M. J. Lazer *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[7] E. Ferrara, "What Types of Covid-19 Conspiracies Are Populated By Twitter Bots?," *First Monday*, vol. 25, no. 6, 2020.

[8] K. Hao, "Nearly half of Twitter accounts pushing to reopen America may be bots," *MIT Technology Review*, May 2020. [Online]. Available: https://www.technologyreview.com/2020/05/21/1002105/covid-bot-twitter-accounts-push-to-reopen-america/.

[9] A. Hern, "Tech giants join with governments to fight Covid misinformation," *The Guardian*, Oct. 20, 2020.

[10] M. D. Kearney, S. C. Chiang, and P. M. Massey, "The Twitter origins and evolution of the COVID-19 'plandemic' conspiracy theory," *Harvard Kennedy*

School Misinformation Review, vol. 1, no. October, pp. 1–18, 2020.

[11] Allcott, H., M. Gentzkow, and C. Yu, “Trends in the diffusion of misinformation on social media”, *Research and Politics 6*(2), 2019.

[12] Twitter, “COVID-19 misleading information policy,” *Twitter Help Center*, 2021. https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.

[13] T. R. Tangherlini, S. Shahsavari, B. Shahbazi, E. Ebrahimzadeh, and V. Roychowdhury, *An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web*, vol. 15, no. 6. 2020.

[14] S. Shahsavari, P. Holur, T. R. Tangherlini, and V. Roychowdhury, “Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news,” *arXiv*, pp. 1–21, 2020.

[15] M. Avram, N. Micallef, S. Patil, and F. Menczer, “Exposure to social engagement metrics increases vulnerability to misinformation,” *Harvard Kennedy School Misinformation Review*, vol. 1, no. 5, pp. 1–11, 2020.

[16] M. Nelimarkka, S.-M. Laaksonen, and B. Semaan, “Social Media Is Polarized”, *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 2018, pp. 957–970.

[17] M. D. Vicario, A. Bessi, F. Zollo, et al., “The spreading of misinformation online”, *Proceedings of the National Academy of Sciences of the United States of America 113*(3), 2016, pp. 554–559.

[18] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The rise of social bots,” *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[19] C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” *Nature Communications*, vol. 9, no. 1, 2018.

[20] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, and S. Stieglitz, “Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks”, *European Journal of Information Systems 28*(4), 2019, pp. 394–412.

[21] E. Noelle-Neumann, “The Theory of Public Opinion: The Concept of the Spiral of Silence”, *Annals of the International Communication Association 14*(1), 1991, pp. 256–287.

[22] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, “Malicious accounts: Dark of the social networks,” *Journal of Network and Computer Applications*, vol. 79, no. November 2016.

[23] D. M. Beskow and K. M. Carley, “Bot conversations are different: Leveraging network metrics for bot detection in Twitter,” *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pp. 825–832, 2018.

[24] S. Kudugunta and E. Ferrara, “Deep neural networks for bot detection,” *Information Sciences*, vol. 467, pp. 312–322, 2018.

[25] C. Grimme, M. Preuss, L. Adam, and H. Trautmann, “Social Bots: Human-Like by Means of Human Control?,” *Big Data*, vol. 5, no. 4, pp. 279–293, 2017,

[26] A. Rauchfleisch and J. Kaiser, “The False positive problem of automatic bot detection in social science research,” *PLoS ONE*, vol. 15, no. 10, 2020.

[27] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “Detection of Novel Social Bots by Ensembles of Specialized Classifiers,” Jun. 2020.

[28] J. Cao, K. A. Basoglu, H. Sheng, and P. B. Lowry, “A systematic review of social networks research in information systems: Building a foundation for exciting future research,” *Communications of the Association for Information Systems*, vol. 36, pp. 727–758, 2015.

[29] I. Himelboim, M. A. Smith, L. Rainie, B. Shneiderman, and C. Espina, “Classifying Twitter Topic-Networks Using Social Network Analysis,” *Social Media and Society*, vol. 3, no. 1, 2017.

[30] I. Anger and C. Kittl, “Measuring influence on Twitter,” in *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, 2011, pp. 4–7,

[31] A. Bovet and H. A. Makse, “Influence of fake news in Twitter during the 2016 US presidential election,” *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[32] B. Huang and K. M. Carley, “Disinformation and Misinformation on Twitter during the Novel Coronavirus Outbreak,” *arXiv*, pp. 1–19, 2020.

[33] A. Patel, “Searching Twitter With Twarc,” *F-Secure blog*, 2018. https://blog.f-secure.com/searching-twitter-with-twarc/.